Conditioning and Stability

Deepayan Sarkar

Condition of a problem

- Abstract problem: compute $f : X \to Y$
- X and Y are normed vector spaces, usually $\mathbb{R}^k$ for some k
- f is referred to as the "problem", and is usually continuous
- We are interested in the behaviour of the problem at a particular "instance" $x \in X$
- A problem instance $f(x)$ is
  - *well-conditioned* if small perturbations in x lead to only small changes in $f(x)$
  - *ill-conditioned* if small perturbations in x can lead to large changes in $f(x)$
- Depending on context, "small" and "large" may be either absolute or relative change

Absolute condition number

- Consider a small perturbation δx in x
- Define the change in f to be $\delta f = f(x + \delta x) - f(x)$
- The *absolute condition number* $\hat{\kappa} = \hat{\kappa}(x)$ of the problem f at x is

$$\hat{\kappa}(x) = \lim_{h \to 0} \sup_{\|\delta x\| \leq h} \frac{\|\delta f\|}{\|\delta x\|}$$

- For readability, this is often written informally as (implicitly assuming δx is infinitesimally small)

$$\hat{\kappa}(x) = \sup_{\delta x} \frac{\|\delta f\|}{\|\delta x\|}$$

- If $f : \mathbb{R} \to \mathbb{R}$ is differentiable, it is easy to see that $\hat{\kappa}(x) = |f'(x)|$
- More generally, if $f : \mathbb{R}^k \to \mathbb{R}$ is differentiable, and $J(x)$ is the Jacobian function, then

$$\hat{\kappa}(x) = \|J(x)\|$$

- Here $\|J(x)\|$ represents a "matrix norm" induced by a vector norm (on $\mathbb{R}^k$)
- Definition: For $A_{m \times n}$, the matrix norm induced by vector norms on $\mathbb{R}^m$ and $\mathbb{R}^n$ is

$$\|A\| = \sup \left\{ \frac{\|Ax\|}{\|x\|} : x \in \mathbb{R}^n, x \neq 0 \right\}$$

- Note that here, the first-order Taylor series expansion of f gives

$$\delta f = f(x + \delta x) - f(x) \approx J(x)\delta x \implies \sup_{\delta x} \frac{\|\delta f\|}{\|\delta x\|} \approx \sup_{\delta x} \frac{\|J(x)\delta x\|}{\|\delta x\|}$$

- Exercise: Show that $\hat{\kappa} = \|J(x)\|$

Relative condition number

- The nature of floating point computations makes it more important to study *relative* changes
- The *relative condition number* $\kappa = \kappa(x)$ of f at x is

$$\kappa(x) = \lim_{h \to 0} \sup_{\|\delta x\| \leq h} \left(\frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right) = \frac{\|x\|}{\|f(x)\|} \lim_{h \to 0} \sup_{\|\delta x\| \leq h} \frac{\|\delta f\|}{\|\delta x\|}$$

- If f is differentiable, we get

$$\kappa(x) = \frac{\|J(x)\| \cdot \|x\|}{\|f(x)\|} = \left| \frac{f'(x)x}{f(x)} \right|$$

- A problem f is well-conditioned if κ is small (e.g., $1, 10, 10^2$) and ill-conditioned if κ is large (e.g., 10^6, $\ldots$)

Examples

- $f(x) = \sqrt{x}, x \geq 0$
  - $f'(x) = \frac{1}{2} x^{-\frac{1}{2}}$
  - So the condition of f at x is
    $$\left| \frac{f'(x)x}{f(x)} \right| = \frac{1}{2} \frac{x^{-\frac{1}{2}}}{x^{\frac{1}{2}}} x = \frac{1}{2}$$
  - So f is well-conditioned for all x.
- $f(x) = x^{\alpha}$
  - Exercise: Condition of f is $|\alpha|$ at all x
- $f(x) = \frac{1}{1-x^2}$
  - $f'(x) = 2x(1-x^2)^{-2}$
  - So condition of f at x is
    $$\left| \frac{f'(x)x}{f(x)} \right| = |2x(1-x^2)^{-2}x(1-x^2)| = \frac{2x^2}{|1-x^2|}$$
  - Can be large for x close to ± 1.
- $f(x_1, x_2) = x_1 - x_2$
  - The Jacobian of f is $J = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 1 & -1 \end{bmatrix}$
  - So $\kappa = \frac{\|J\| \cdot \|x\|}{|x_1 - x_2|}$
  - What is $\|x\|$? Common choices are
    - L_1: $|x_1| + |x_2|$
    - L_2: $\sqrt{x_1^2 + x_2^2}$
    - L_∞: $\max\{|x_1|, |x_2|\}$
  - What is $\|J\|$? Depends on vector norm, but some constant c for this J regardless of choice

- So κ is, with the L_∞ norm, $\kappa = \frac{c \max\{|x_1|, |x_2|\}}{|x_1 - x_2|}$
- Ill-conditioned when $x_1 \approx x_2$

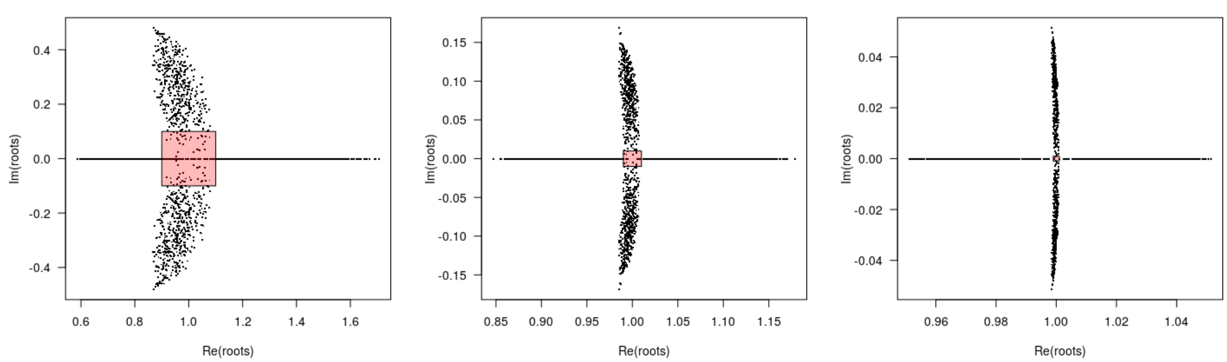
- Roots of polynomials: e.g., $ax^2 + bx + c = 0$

$$f(a, b, c) = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

- Exercise: Show that for $x^2 - 2x + 1 = (x-1)^2 = 0$, $f(1, -2, 1)$ has $\kappa = \infty$
- Hint: try perturbing one coefficient at a time
- Graphical demonstration:

```
qroot <- function(coefs) {
    a <- coefs[1]; b <- coefs[2]; c <- coefs[3]
    C <- sqrt(complex(real = b^2 - 4 * a * c, imaginary = 0))
    (-b + c(-1, 1) * C) / (2 * a)
}
abc <- c(1, -2, 1)
```

```
par(mfrow = c(1, 3))
for (eps in c(0.1, 0.01, 0.001))
{
    roots <- replicate(1000, qroot(abc + eps * runif(3, -1, 1)))
    plot(roots, pch = ".", cex = 3, las = 1)
    rect(1-eps, -eps, 1+eps, eps, col = "#FF000044")
}
```



Formal model for floating point arithmetic

- Recall that floating point numbers are represented as

$$\text{significand} \times \text{base}^{\text{exponent}}$$

- Ignoring the limitations imposed by the finite range of the exponent, define

$$\mathbb{F} = \{0\} \cup \left\{ \pm \frac{m}{2^t} \times 2^e : e \in \mathbb{Z} \text{ and } m \text{ integer with } 1 \leq m \leq 2^t \right\}$$

- Here the integer t is the *precision* of the representation (usually 24 or 53)
- e can be an arbitrary integer, so there is no "overflow" or "underflow" ($\mathbb{F} = 2\mathbb{F}$)

- This is still a useful formal model for the subset of $\mathbb{R}$ that has a floating point representation
- For example, with $t = 53$,

$$\begin{aligned}
\mathbb{F} \cap [1,2] &= \{1, 1+2^{-52}, 1+2\times 2^{-52}, 1+3\times 2^{-52}, ..., 2\}, \\
\mathbb{F} \cap [2,4] &= \{2, 2+2^{-51}, 2+2\times 2^{-51}, 2+3\times 2^{-51}, ..., 4\}, \text{etc.}
\end{aligned}$$

Machine epsilon

- The resolution of $\mathbb{F}$ is quantified by a number known as *machine epsilon*, ϵ_m
- Let us tentatively define ϵ_m to be half the distance between 1 and the next larger number in $\mathbb{F}$
- Clearly, $\epsilon_m = \frac{1}{2} \times 0.000\cdots 0001 = \frac{1}{2} \times 2^{t-1} = 2^{-t}$, and has the following property:

  For all $x \in \mathbb{R}$, there exists $x^* \in \mathbb{F}$ such that $|x - x^*| \leq \epsilon_m \cdot |x|$
- For $t = 24$ (Float32), $\epsilon_m = 2^{-24} \approx 6 \times 10^{-8}$
- For $t = 53$ (Float64), $\epsilon_m = 2^{-53} \approx 1.1 \times 10^{-16}$
- For any $x \in \mathbb{R}$, define $\text{fl}(x)$ to be the element in $\mathbb{F}$ closest to x
- Then, a restatement of the above property is

  For all $x \in \mathbb{R}$, there exists ϵ with $|\epsilon| \leq \epsilon_m$ such that $\text{fl}(x) = x(1+\epsilon)$
- In other words, the relative *approximation error* of any real number is bounded by ϵ_m

Arithmetic of floating point numbers

- Consider the elementary arithmetic operations $+, -, \times, \div$
- How should we expect these to behave on $\mathbb{F}$?
- Let $*$ denote one of these elementary operations, and $\circledast$ denote the corresponding operation on $\mathbb{F}$
- Then we would ideally want, for $x, y \in \mathbb{F}$,

$$x \circledast y = \text{fl}(x * y)$$

- If this is indeed true, then we have the **Fundamental axiom of floating point arithmetic**:

  For all $x, y \in \mathbb{F}$, there exists ϵ with $|\epsilon| \leq \epsilon_m$ such that $x \circledast y = (x * y)(1+\epsilon)$
- In practice, this may not hold for the theoretical ϵ_m, but only for some larger value
- The smallest ϵ_m for which this is guaranteed (on a given machine) is defined to be the machine epsilon

Algorithms and stability

- Suppose we want to solve a problem $f : X \to Y$
- There can be multiple *algorithms* to calculate a candidate solution
- Let $\tilde{f} : X \to Y$ be the actual implementation of an algorithm to solve f
- At a minimum, this will involve the approximation of x by $\text{fl}(x)$
- In practice, suppose we want to calculate $f(x)$, and actually compute $\tilde{f}(x)$
- The relative error is

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|}$$

- Recall that $\mathrm{fl}(x) \approx x(1+\epsilon_m) \implies \frac{\|\mathrm{fl}(x)-x\|}{\|x\|} \approx \epsilon_m$
- If $\kappa = \kappa(x)$ is the relative condition number of $f(x)$, we expect (note: for f, not $\tilde{f}$)

$$\frac{\|f(\mathrm{fl}(x)) - f(x)\|}{\|f(x)\|} \approx \kappa \frac{\|\mathrm{fl}(x) - x\|}{\|x\|} \approx \kappa\epsilon_m$$

- This is the best we can hope for with $\tilde{f}$ instead of f

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} \approx \kappa\epsilon_m$$

- Informally, an algorithm $\tilde{f}$ is *unstable* if this does not hold

Instability

- Instability arises due to ill-conditioned intermediate steps in an algorithm $\tilde{f}$
- The basic idea is to compare the (inherent) condition of $f(x)$ with the conditions of intermediate steps
- Badly conditioned intermediate steps make the process unstable.

Instability: a toy example

- To make the idea concrete, consider the problem: $f(x) = \sqrt{x+1} - \sqrt{x}, x > 0$
- It is easily seen that the condition of f at x is $\frac{1}{2}\frac{x}{\sqrt{x+1}\sqrt{x}} \approx \frac{1}{2}$ when x is large
- A possible algorithm $\tilde{f}$, directly using the definition, will proceed as follows
  - $x_0 = x$
  - $x_1 = x_0 + 1$
  - $x_2 = \sqrt{x_1}$
  - $x_3 = \sqrt{x_0}$
  - $x_4 = x_2 - x_3$
- In general, suppose $y = \tilde{f}(x)$ is computed in n steps
- Let x_i be the output of the ith step (define $x_0 = x$)
- Then $y = \tilde{f}(x) = x_n$ can also be viewed as a function of each of the intermediate x_is
- Denote the ith such function by $\tilde{f}_i$, such that $y = \tilde{f}_i(x_i)$
- In particular, $\tilde{f}_0 = \tilde{f}$
- Then the instability in the total computation is dominated by the most ill-conditioned $\tilde{f}_i$
- For the $\tilde{f}$ given above, we have
  - $\tilde{f}(t) = \sqrt{t+1} - \sqrt{t}$
  - $x_0 = x \implies \tilde{f}_0(t) = \sqrt{t+1} - \sqrt{t}$
  - $x_1 = x_0 + 1 \implies \tilde{f}_1(t) = \sqrt{t} - \sqrt{x_0}$
  - $x_2 = \sqrt{x_1} \implies \tilde{f}_2(t) = t - \sqrt{x_0}$

  - $x_3 = \sqrt{x_0} \implies \tilde{f}_3(t) = x_2 - t$

- Consider the condition of $\tilde{f}_3 = x_2 - t$, which is (treating x_2 as fixed)

$$\left| \frac{\tilde{f_3}'(t)\, t}{\tilde{f}_3(t)} \right| = \left| \frac{t}{x_2 - t} \right|$$

- This can be arbitrarily large for large x, e.g.,

```
x <- c(10, 100, 1000, 10000); t <- sqrt(x)
abs(t / (sqrt(x+1) - t))
```

```
[1]    20.48809   200.49876  2000.49988 20000.49999
```

- Here x_2 and t are related, but the condition number is w.r.t. perturbations in t keeping x_2 fixed
- An alternative formula for f is $f(x) = \frac{1}{\sqrt{x+1}+\sqrt{x}}$
- An algorithm based on this formula would proceed as
  - $x_0 = x \implies \tilde{f}_0(t) = \frac{1}{\sqrt{t+1}+\sqrt{t}}$
  - $x_1 = x_0 + 1 \implies \tilde{f}_1(t) = \frac{1}{\sqrt{t}+\sqrt{x_0}}$
  - $x_2 = \sqrt{x_1} \implies \tilde{f}_2(t) = \frac{1}{t+\sqrt{x_0}}$
  - $x_3 = \sqrt{x_0} \implies \tilde{f}_3(t) = \frac{1}{x_2+t}$
  - $x_4 = x_2 + x_3 \implies \tilde{f}_4(t) = \frac{1}{t}$
  - $x_5 = 1/x_4$
- Exercise: All these have good condition when t is large